The PAHN-PaN NFDI Consortium

Correspondence: t.kollegger@gsi.de, thomas.schoerner@desy.de, andreas.haungs@kit.edu

1 BINDING LETTER OF INTENT

This is the **binding Letter of Intent** of the PAHN-PaN NFDI Consortium.

2 FORMAL DETAILS

Planned name of consortium
Particle, Astroparticle, Hadron and Nuclear Physics accelerates the NFDI

Acronym of the planned consortium
PAHN-PaN

Applicant institution
Deutsches Elektronen-Synchrotron (DESY), Notkestr. 85, D-22607 Hamburg
Prof. Dr. Joachim Mnich, *joachim.mnich@desy.de*, Director for Particle Physics

Spokesperson
Priv.-Doz. Dr. Thomas Schörner, *thomas.schoerner@desy.de*, DESY

Co-applicant institutions
— Albert-Ludwigs-Universität Freiburg (Prof. Dr. Markus Schumacher, *markus.schumacher@physik.uni-freiburg.de*, Physikalisches Institut, Hermann-Herder-Str.3, D-79104 Freiburg)
— Bergische Universität Wuppertal (Prof. Dr. Christian Zeitnitz, *zeitnitz@uni-wuppertal.de*, Fachbereich C — Physik, Gaußstrasse 20, D-42119 Wuppertal)
— Frankfurt Institute for Advanced Studies (Prof. Dr. Volker Lindenstruth, *voli@compeng.de*, Ruth-Moufang-Straße 1, D-60438 Frankfurt am Main)
— Friedrich-Alexander-Universität Erlangen-Nürnberg (Prof. Dr. Ulrich Katz, *uli.katz@physik.uni-erlangen.de*, Physikalisches Institut, Lehrstuhl für Experimentalphysik, Erwin-Rommel-Str. 1, D-91058 Erlangen)
— Georg-August-Universität Göttingen (Prof. Dr. Arnulf Quadt, *aquadt@uni-goettingen.de*, Friedrich-Hund-Platz 1, D-37077 Göttingen)
— GSI Helmholtz-Zentrum für Schwerionenforschung GmbH (Dr. Thorsten Kollegger, *t.kollegger@gsi.de*, FAIR GmbH, Planckstr. 1, D-64291 Darmstadt)
— Johannes-Gutenberg-Universität Mainz (Prof. Dr. Volker Büscher, *buescher@uni-mainz.de*, ETAP, Staudingerweg 7, D-55128 Mainz)
— Karlsruher Institut für Technologie (Dr. Andreas Haungs, *andreas.haungs@kit edu*, Institut für Kernphysik, Campus Nord, Geb. 401, Postfach 3640, D-76021 Karlsruhe)
— Ludwig-Maximilians-Universität München (Prof. Dr. Thomas Kuhr, *Thomas.Kuhr@lmu.de*, Fakultät für Physik, Excellence Cluster Universe. Boltzmannstr. 2, D-85748 Garching)
— Rheinische Friedrich-Wilhelms-Universität Bonn (Priv.-Doz. Dr. Philip Bechtle, *bechtle@physik.uni-bonn.de*, Physikalisches Institut, Nussallee 12, D-53115 Bonn)
— Ruprecht-Karls-Universität Heidelberg (Prof. Dr. Tilman Plehn, *plehn@uni-heidelberg.de*, Institut für Theoretische Physik, Philosophenweg 16, D-69120 Heidelberg)

— RWTH Aachen University (Prof. Dr. Alexander Schmidt, *alexander.schmidt@physik.rwth-aachen.de*, Lehrstuhl für Experimentalphysik IIIA, Sommerfeldstraße 16, D-52074 Aachen)
— Technische Universität Darmstadt (Dr. Stefan Typel, *s.typel@gsi.de*, Institut für Kernphysik, Schlossgartenstraße 9, D-64289 Darmstadt)
— Technische Universität Dortmund (Prof. Dr. Kevin Kröninger, *kevin.kroeninger@cern.ch*, Experimentelle Physik IV, D-44221 Dortmund)
— Universität Bielefeld (Prof. Dr. Frithjof Karsch, *karsch@physik.uni-bielefeld.de* , Fakultät für Physik, Universitätsstr. 25, D-33615 Bielefeld)
— Universität Hamburg (Jun.-Prof. Dr. Gregor Kasieczka, *gregor.kasieczka@desy.de*, Institut für Experimentalphysik, Luruper Chaussee 149, D-22761 Hamburg)
— Universität zu Köln (Dr. Jan Mayer, *jmayer@ikp.uni-koeln.de*, Institut für Kernphysik, Zülpicher Straße 77, D-50937 Köln)
— Universität Regensburg (Prof. Dr. Gunnar Bali, *gunnar.bali@physik.uni-r.de*, Institut für Theoretische Physik, Fakultät für Physik, D-93040 Regensburg)
— Westfälische Wilhelms-Universität Münster (Dr. Raimund Vogl, *rvogl@uni-muenster.de*, Zentrum für Informationsverarbeitung, Röntgenstraße 7-13, D-48149 Münster)
— Technische Universität München (Prof. Dr. Nora Brambilla, *nora.brambilla@ph.tum.de*, Fakultät für Physik, James-Franck-Str. 1, D-85748 Garching)

Participants
CERN - European Organization for Nuclear Research (Dr. Markus Elsing)
Deutsche Physikalische Gesellschaft (Dr. Georg Düchs)
Friedrich-Schiller-Universität Jena (Prof. Dr. Stefan Fritzsche)
Helmholtz-Institut Jena (Prof. Dr. Thomas Stöhlker)
Helmholtz-Zentrum Dresden Rossendorf (Dr. Michael Bussmann)
Johann Wolfgang Goethe-Universität Frankfurt (Prof. Dr. Hannah Elfner)
Julius-Maximilians-Universität Würzburg (Prof. Dr. Thomas Trefzger)
Justus-Liebig-Universität Giessen (Prof. Dr. Lorenz von Smekal)
Max-Planck-Institut für Physik München (Dr. Stefan Kluth)
Max-Planck-Institut für Kernphysik Heidelberg (Prof. Dr. Michael Schmelling)
Ruhr-Universität Bochum (Prof. Dr. Ulrich Wiedner)
Technische Universität Braunschweig (Prof. Dr. Andrey Surzhykov)
Technische Universität Dresden (Prof. Dr. Arno Straessner)
TIB — Leibniz Information Center for Science and Technology (Prof. Dr. Sören Auer)
Universität Siegen (Prof. Dr. Ivor Fleck)

3 OBJECTIVES, WORK PROGRAMME AND RESEARCH ENVIRONMENT

3.1 Research area of the proposed consortium

Primary: 309-01 Nuclear and Elementary Particle Physics, Quantum Mechanics, Relativity, Fields
Secondary: 311-01 Astrophysics and Astronomy
Secondary: 312-01 Mathematics
Secondary: 409 Computer Science

3.2 Summary of the planned consortium's main objectives and task areas

The German research communities in the fields of particle, astroparticle, and hadron & nuclear physics (in short "PAHN") have a long standing tradition of cooperation. The respective committees (KET, KAT and KHuK) represent more than 3000 scientists with doctoral degree. The research is typically carried out in large international collaborations and at large-scale research facilities. Existing structures like the Helmholtz alliances Terascale, EMMI, and HAP, as well as coordinated actions like the German inputs to the update of the European Strategy for Particle Physics, suggest a joint NFDI effort.

Driven by the needs created by their research, the communities involved in the PAHN-PaN Consortium have always been at the forefront of technological developments. Currently, due to the development of new accelerators, new observatories and experiments, and new detectors with increased resolutions and higher event rates, the PAHN physics is experiencing a rapid increase of data rates and data volumes. This boost of data leads to ever increasing demands on data analysis power and data management capabilities and requires new methods in data analytics. As an example, for the High-Luminosity LHC, storage

needs up to ten times higher than at today's LHC are predicted. Also other upcoming facilities, like the FAIR accelerator complex[1] or the next generation of observatories in astroparticle physics like CTA, KM3NeT, IceCube or the upgraded Pierre Auger Observatory will provide data amounts in the ExaByte regime already in the 2020s.

The challenge of an efficient handling of such data volumes cannot be met by today's approaches. Rather, new ideas, methods, concepts, and strategies are required to harvest, manage, access, and analyse the data, and to publish the results, as well as to offer open and public access of the research data; i.e. to provide a life cycle of scientific data in agreement with the *FAIR* guiding principles for scientific data management and stewardship. This challenge urges the PAHN communities to once again become drivers of technological developments, spearheading "big data" analytics for the entire scientific community and making their experience and expertise available for and through the NFDI.

The goal of the PAHN-PaN Consortium is to prepare the involved communities for the challenges by developing solutions for the problems that they are experiencing, and to help setting up the necessary structures in order to achieve that. These structures will be formed in a way to allow exploiting synergies within the consortium; easy connection and transfer of knowledge and technology to and from neighbouring consortia and communities; and the establishing of relevant services for PAHN-PaN and other consortia. More concretely, PAHN-PaN's main objectives can be summarised in the following directions of activities: the development of interdisciplinary cloud-enabled data workflows for an efficient use of heterogeneous and federated infrastructures and environments, their application in generalised and standardised large-scale data management infrastructures following the *FAIR* data principles, the ensuing development of sustainable software solutions and analysis methods, and training and education of scientists from the PAHN-PaN consortium and others.

The majority of PAHN activities are carried out in large international collaborations working at large-scale facilities. The goal of PAHN-PaN to develop and structure the management of the full data life cycle of these facilities can certainly only be achieved within the context of the international collaborations and in close cooperation with them. With the PAHN-PaN consortium, however, Germany will be a lead player and forerunner in the efforts for a global data management plan.

Considering the full data life cycle from generation of the data to the public (re-) use, PAHN-PaN's goal for the next decade is not only to develop the content of a data management infrastructure, but also to provide a standardised handbook describing the tools, methods and services for an efficient data management in all steps of the life cycle (collecting, processing, analysing, storing, sharing, finding). This handbook shall be used as guide for the PAHN scientists at existing and future facilities, but also as a blueprint for other science fields. In PAHN-PaN, the scientists will focus on the development of data management related software and tools for an optimised operation of (federated) computing and data infrastructures. The typical workflows also involve Monte Carlo event generators or simulations that are developed and need to be available as open source codes for the experimental collaborations but also to theoretical groups. A standardised way to compare such simulations to experimental data with the goal of quality control of such simulations, applying the *FAIR* principles, is important.

The science culture in the PAHN fields is in some aspects similar, but nonetheless different from other research fields, like e.g. astronomy or material sciences. This is true for particle and hadron & nuclear physics. However, the large astroparticle physics observatories are in a hybrid phase: the facilities work with particle physics detection methods, the reconstructed final data are of astrophysical interest. Therefore, the KAT community is involved in both PAHN-PaN and ASTRO-NFDI.

Close co-operations will be pursued wherever a mutual benefit is conceivable. Initially, PAHN-PaN will act in close collaboration with e.g. ASTRO-NFDI and the DAPHNE Consortium (photon and neutron science). While the fields of astronomy and PAHN have ample experience in dealing with "big data", it is expected that photon and neutron experiments will soon require similar skills and new concepts for data reduction and compression. The three NFDI consortia ASTRO-NFDI, DAPHNE and PAHN-PaN will therefore combine forces and use their expertise and versatility for the entire science system. Immediate collaboration is also envisaged with the field of mathematics, where for example statistical procedures in big data analytics will be explored together with the MaRDI Consortium. In addition, via the DPG (German Physics Society) there will be tight connection to the entire research field of physics in order to generalise and standardise the *FAIR* life cycle of scientific data.

All technical efforts in PAHN-PaN will be accompanied by outreach and knowledge transfer activities. Furthermoe, cross-cutting topics, like the standardisation of data curation procedures for open data including legal issues of data ownership, will be pursued that facilitate and profit from collaboration and exchange with other consortia.

The above considerations lead us to the following specific task areas, which are described in more detail in section 5.2 in the annex:

(1) Developing workflows and tools for data management;

(2) *FAIR* data management infrastructures and open data;

[1]The acronym "FAIR" is used in this document with two different meanings: On the one hand the accelerator complex "Facility for Antiproton and Ion Research" at GSI Darmstadt, and on the other hand the principles of data management (https://www.force11.org/fairprinciples). For differentiation we will always speak of 'FAIR accelerator' in the case of the accelerator facility. The *FAIR* data principles will always be set in italics.

(3) Data analysis procedures and services;

(4) Online analysis and data reduction.

They are complemented by the cross-cutting topics described in section 4:

(A) Synergies, knowledge transfer and collaboration;

(B) Services;

(C) Training, education and outreach,

and an efficient governance structure, described in section 5.3.

3.3 Proposed use of existing infrastructures, tools and services

Data sources for the PAHN communities are usually experiments and instruments built and operated in large international collaborations: Germany is strongly involved in the LHC experiments at CERN and in Belle II at KEK. Large instruments are also being constructed at the FAIR accelerator facility at GSI in Darmstadt. Other examples are observatories like Auger, CTA, IceCube and KM3NeT. These large experiments are complemented with smaller and dedicated ones like the experiments at CERN's SPS, the KATRIN neutrino experiment at KIT, or experiments that are installed at national accelerators like MAMI or the future MESA and S-DALINAC facilities. Theoretical physics results from super-computer calculations also lead to sizable data volumes. Many experiments in the field create data volumes of a size that makes it unfeasible to host them at a single facility. Consequently, and already many years ago, the involved communities developed tools to transparently interconnect various sites. The prime example is the Worldwide LHC Computing Grid (WLCG) that spans over 170 computing sites world wide. Germany contributes roughly 10% of the total WLCG capacity. The Helmholtz centers DESY, GSI and KIT, Max-Planck institutes and various universities are committed to support WLCG and have thus a long-standing experience in large-scale data management and processing. The involved technology as well as the existing infrastructure is also used by experiments beyond the LHC community. By design the resources are shared among all members of an experiment.

The internationally distributed infrastructure is typically used for large and demanding processing in batch mode. In order to provide a convenient environment for more interactive analysis, many countries host national analysis facilities. In Germany, DESY and KIT provide the National Analysis Facility (NAF) and the Tier-1 center GridKa, respectively, for all major German particle physics groups. A similar facility for the German hadron & nuclear physcics community is provided at GSI. These facilities are built based on state of the art hardware including GPUs.

With the tremendously growing data rates at the experimental facilities of the PAHN communities, there is a significantly growing demand for a large-scale federated computing infrastructure, which in turn requires an elaborated research data management. The central question resulting from this is the requirement for a reasonably priced, sustainable and efficient compute infrastructure that satisfies the demand and that can help to develop future-oriented concepts. It is particularly important to further develop the corresponding experiment software, and also the development of new and innovative data management software will be necessary. Development of data reduction, data compression and data filtering methods to reduce the data volumes is equally important. The information infrastructure in Germany within this research field is already outstanding. However, the current compute infrastructure needs to be transformed such that real-time online analysis applying new and efficient algorithms is possible. Modern, distributed computing workflows with user-friendly remote access to distributed storage and compute resources will be based on user portals for data analysis, visualisation, control and simulation.

Moreover, for the future it is necessary to provide an efficient integration of so-called opportunistic resources (e.g. scientific and commercial cloud providers, larger HPC centres as well as models like volunteer computing). One more important point is the concentration to few large centres storing and providing experimental data that together can be seen as a single virtual data centre. Such a data lake needs to be equally accessible via all state-of-the-art access methods and from all participating compute resources, which therefore need to be connected with sufficiently high bandwidth. Efficient high bandwidth data transport is paramount to this effort. Building on existing experience in technologies (dCache, XRootD) PAHN-PaN will work on next-generation federated storage infrastructures. Complementing the infrastructure for data storage, the infrastructure for metadata and source code management also needs to be developed to withstand the data deluge. This is in line with the adoption of *FAIR* principles and — where possible — open science. In order to facilitate the development of new and efficient computing infrastructures, parts of the existing computing infrastructure will be provided as test beds.

The communities of the PAHN-PaN consortium are either developers or experienced large-scale users of the middleware components that are employed to set up the mentioned structures. Most of the involved base tools for distributed computing, multi-PetaByte storage solution and services to transfer data between them are not domain-specific and are therefore applicable elsewhere.

These ideas are in line with the ”Digitale Agenda” of the German federal government, and they are exceeding the initiatives within ErUM-Data. The general overarching idea of all suggestions is a strengthening of science. The further development of infrastructures will also further strengthen Germany’s position in the international scientific landscape. Additionally, the participating communities of the consortium already today provide important contributions to the development of technology and to the education of qualified personnel — an aspect of decisive importance for Germany as a competitive industry location. The broad technical competence, the diverse and interdisciplinary blend of physicists, engineers, and technicians, together with a rich technical laboratory infrastructure, as it is provided by PAHN-PaN, are key assets in achieving the consortium’s goals.

3.4 Interfaces to other proposed NFDI consortia

- The communities from the ASTRO-NFDI, DAPHNE and PAHN-PaN consortia already have a link via the BMBF’s ErUM-Data initiative. While there is still the need for domain-specific solutions that satisfy the complex demands of individual research areas, there are certain areas that can be addressed in common. Within ErUM-Data, a cross-community platform ”partnership for digitalization” has been proposed. This partnership will foster the collaboration among the communities. Similar links are foreseen also between the various NFDI consortia.
- Close collaboration with the astronomy consortium ASTRO-NFDI is of significant importance: Different as our concrete situations and experiences may be, both consortia are facing heavy increases in data volumes and thus very similar problems. Due to the size of the two communities, compared to many other players in the NFDI, there is a certain responsibility to combine forces and use the expertise and versatility for the entire science system. Identified topics are big data management, data reduction, and open data. As pointed out above, astroparticle physics, participating in both consortia, plays a specific role.
- Immediate collaboration is also envisaged with the field of mathematics (i.e. the consortium MaRDI, the Mathematical Research Data Initiative), where e.g. common concepts of data integration and annotation with metadata will be developed, and where viable analysis methods or statistical procedures including machine learning are explored together, using the mathematicians’ specific expertise and our wealth in experimental data.
- PAHN-PaN, ASTRO-NFDI, MaRDI and other consortia plan to cooperate with the cross-section working group ”Software & Online applications, Software Curation”. One important aspect to follow in this task is based on the fact that sustainability of software never comes from the software itself but only through continuous and long-term development, maintenance and support of the codes by competent scientific software experts.
- Via the DPG (German Physics Society) there will be tight connection to the entire research field of physics, with the aim of generalising and standardising the *FAIR* life cycle of scientific data. The DPG may also handle cross-cutting topics and synergy effects among physics NFDI consortia. Examples are workshops on topics of common interest as well as education and training.

4 CROSS-CUTTING TOPICS

In order to foster cross-consortium synergies, services and tools developed for research field specific applications in the corresponding task areas need to be generalised and adapted to be useful for the entire NFDI. In order to foster this concept, we define the following tasks as cross-cutting topics:

Cross-cutting topic (A) ”Synergies, knowledge transfer and collaboration”: Here, we will specifically address cross-NFDI topics and try to exploit the interaction with other consortia on technical matters, to our mutual benefit. Also interaction with industry will be handled here. As first concrete examples, the following links will initially be pursued:

- Similar to many other sciences, PAHN-PaN scientists work on advanced machine learning (ML) algorithms for data analysis (GANs, Auto-encoder, ...) and more generally also on the application of new ML developments. Again, we expect strong collaboration with astronomy, mathematics etc. and also industry. One aspect in this respect are ML training data sets from particle, astroparticle, hadron & nuclear physics that could be exploited by other NFDI consortia in order to develop new methods and algorithms. Also statistical analysis procedures may be explored on corresponding training data sets.
- Similarly, big data tools are a field that promises synergies between various consortia. Examples are the Kubernetes Cluster on Google Cloud or native data management software like dCache, XRootD, Dynafed, or RUCIO, as well as transfer services like FTS3 that are frequently used within PAHN. Important related cross-cutting topics are data reduction and open data.

- Dealing with research software is a topic of high relevance for many NFDI consortia. Spheres of activity are the sustainability of research software, the introduction of standards in scientific software development, software publication, and proper career paths for developers.
- Synergies exist also in the field of training and education, and maybe also outreach (see also cross-cutting topic C), where the forces of several consortia may be bundled to train the next generation of data scientists.
- The PAHN community traditionally uses also resources and infrastructure provided by the large HPC centers in Germany. Many HPC centers extend their activities and services towards ML, big data and cloud computing. For PAHN-PaN, we plan close cooperation with these initiatives that could provide a cornerstone for the required hardware infrastructure and serve as competence centers for new hardware architectures.

Further collaborations will be explored in the course of the PAHN-PaN project.

Cross-cutting topic (B) "Services": The PAHN community has development and operations experiences with large-scale distributed services for the processing and management of scientific data. Many services are composed of a layered structure, where base services are not specific to a particular experiment and can therefore also be used in other domains. Members of the community are actively involved in national and international programmes that aim at providing generic services for scientific computing, including Monte Carlo simulation methods, machine learning and analysis application procedures. There are ongoing efforts to adopt services (like those provided in the EOSC) to concrete applications that arise from the needs of the experiments. Such experiences can serve as blue prints for other communities that are dealing with similar challenges. The effort spent on this topic is mainly targeting at providing and supporting generic solutions.

A data analysis tool that has become indispensible for scientists of all disciplines is machine learning (ML). Applying ML in PAHN's research fields is especially attractive because of the availability of high-quality labeled data from simulations in large-scale and diverse collaborations. As part of the NFDI, PAHN-PaN will build on this by developing a service that provides automated ML for scientific data in the PAHN domain.

Another concrete example for cross-cutting services may be further developed and offered by TIB, the Leibniz Information Center for Science and Technology. TIB has long-time expertise in curating metadata and in developing standardised and inter-operable metadata schemes tailored to special use cases: research data services like DataCite and ORCID; the interlinking of articles and data; or the semi-automatic semantic annotation of data using ML techniques.

The Cross-Community Collaboration Board (see below) will be instrumental in providing the relevant contacts to other communities and consortia, and in communicating both their needs and boundary conditions with respect to services developed or offered by PAHN-PaN. The exchange between different consortia in such a body will also help to target deeper and deeper service layers and thus to make better use of infrastructures and effort. It is understood that the implementation of services across community and consortium boundaries will also lead to (currently still unforeseeable) structures linking the various players.

Cross-cutting topic (C) "Training, education, and outreach": The communities involved in PAHN-PaN have a long-standing history in organising education and training events, not least from the times of the above-mentioned Helmholtz alliances in which education and training of the entire community were major work packages. A focus always was and will be in future on enabling a large part of our communities to fully exploit the power and the options of the provided national research infrastructures.

For PAHN-PaN, we foresee a bundling and streamlining of our training and education programmes across all consortium partners, and — more importantly — we will open our programmes for other consortia. We foresee events for scientists at all career levels — from the master and Ph.D. student level to senior scientists — ranging from lecture-style workshops over hands-on tutorials to developers' meetings during which concrete progress on certain problems is achieved. We will also support the education of future data stewards and data scientists. The training and education programme will be designed to be complete and consistent, and it will focus on the most relevant future technologies in computing, data management, etc.

5 ANNEX

5.1 Information on spokesperson and applicant institution

Applicant institution: Deutsches Elektronen-Synchrotron (DESY)
Notkestr. 85, D-22607 Hamburg

Head of institution: Prof. Dr. Joachim Mnich, Director for Particle Physics

joachim.mnich@desy.de, phone +49 40 8998 1921

PAHN-PaN Spokesperson: Priv.-Doz. Dr. Thomas Schörner
thomas.schoerner@desy.de, phone +49 40 8998 3429

Scientific degrees: Dr. rer. nat., Ludwig Maximilians University Munich (2001)
Habilitation, Hamburg University (2009)

Current position: Senior scientist at DESY (since 2008)

Previous positions: Scientific assistant, Hamburg University (2003-08)
Research Fellow, CERN, Geneva (2001-2003)
Post-doc, MPI for Physics Munich (2001)

Other functions: Scientific Manager / Leader of the Analysis Center,
Helmholtz Alliance "Physics at the Terascale" (2008-14)
Project manager, E-JADE MCSR Action under Horizon-2020
Co-organiser of several international schools in the field of particle physics

Collaborations in the past 3 years:
Prof. Philip Burrows, Oxford University, Oxford, UK
Dr. Nick Ellis, CERN, Geneva, Switzerland
Prof. Dr. Lutz Feld, RWTH Aachen University, Aachen, Germany
Prof. Dr. Keisuke Fujii, KEK, Tsukuba, Japan
Prof. Dr. Juan Fuster, CSIC, University of Valencia, Valencia, Spain
Prof. Dr. Erika Garutti, Hamburg University, Hamburg, Germany
Prof. Dr. Thomas Hebbeker, RWTH Aachen University, Aachen, Germany
Dr. Andreas Heiss, KIT, Karlsruhe, Germany
Prof. Dr. Ulrich Husemann, KIT, Karlsruhe, Germany
Prof. Dr. Peter Schleper, Hamburg University, Hamburg, Germany
Prof. Dr. Steinar Stapnes, CERN, Geneva, Switzerland
Dr. Yasuhiro Sugimoto, KEK, Tsukuba, Japan
Prof. Dr. Maxim Titov, CEA Saclay, Gif sur Yvette Cedex, France
Dr. Roman Poeschl, LAL Orsay, Universite Paris-Sud, Orsay CEDEX, France

5.2 Task areas in detail

Task area 1 "Developing workflows and tools for data management"

This task area defines and develops data handling standards and data processing workflows which are as generic and interoperable as possible while respecting the *FAIR* data principles. In this context, data can be either raw experimental data or data that comprise all information on the experimental apparatus or codes in theoretical calculations/simulations that were used to generate data. High-level services necessary to implement the data processing workflows will be selected and (further) developed. This includes existing data management software like dCache, XRootD, Dynafed, RUCIO, or IRODS, as well as transfer services like FTS3. Workflows will be based on the services of the task area "*FAIR* data management infrastructures and open data", which will build the foundation of the distributed computing and data management environment. There is an increasing demand for the ability to utilize a spectrum of resources including HPCs and cloud systems. In addition, upcoming workflow tools need to support specialised hardware architectures like GPUs and FPGAs. In particular, the following items will be pursued: workflows and middleware of specific and generic data access methods including authentification and authorisation, data security and access rights; workflows and middleware to generate standardised (cross-disciplinary) meta-data; user-transparent inclusion of heterogenous, opportunistic and long-term IT resources into data processing workflows; workflows and middleware to support the definition of application-specific machine learning architectures. It is important to us that this work is carried out in accordance with the international activities and collaborations in the research field. Furthermore, services, standards and solutions developed in this task area also need to fit into global structures such as the European Open Science Cloud (EOSC).

Task area 2 "*FAIR* data management infrastructures and open data"

The data of the PAHN-PaN facilities contain a high potential for added value, especially for cross-experiment, cross-theory, and even cross-disciplinary usages. The main idea of this task is to apply the data handling standards and data processing workflows developed in task area 1 and to build full-scale data management infrastructure prototypes based on selected high-level data

management services. This will not only address the requirements of the experiments but also of accompanying large-scale theoretical computer simulations (e.g. lattice QCD, simulations in nuclear astrophysics and in astroparticle physics, or of heavy-ion collisions). In this context the concept of data lakes plays an important role. These infrastructures need to be scaled up to production level so that they can be used productively in all participating communities. In order to ensure inter-operability with other communities, it is also important that these data management infrastructures make use of common standards whenever possible. In addition, the further development of cross-disciplinary standards for data handling, curation, preservation, and data publication needs to be pursued. A typical application example is the multi-messenger ansatz currently pursued in astroparticle physics.

The long-term usability of scientific data requires sophisticated storage methods that allow for fast and sustained access as well as for mining the data on request. In order to achieve this, data management and stewardship have to be fostered; specialists are needed who support the science data management in their communities and who develop the data management procedures (including the full data life cycle as well as generation and curation of metadata, the generalization of data models, and the development of standardized and interoperable metadata schemes tailored to special use cases). First cross-experimental infrastructures already exist, e.g. the DPHEP at CERN or the data center KCDC for astroparticle physics. For future challenges in big data analytics, however, a coherent concept has to be developed. These developments will benefit from the collaboration with computer scientists, software developers, mathematicians, scientific librarians, and scientists from other fields.

Task area 3 ”Data analysis procedures and services”

A data infrastructure only has a scientific value if tools exist to analyse the data. The exceedingly large data sets require properly modified big data methods because the typical workflow of analyzing data on a local computer in an interactive way cannot be simply scaled to data distributed on heterogeneous resources. Algorithms must be adjusted to work on partitioned data and to efficiently run on various hardware architectures. With the PAHN expertise in a broad spectrum of applications ranging from lattice QCD calculations to experiment-specific reconstruction algorithms, we will investigate if new technological developments can be exploited to augment or replace community-specific solutions. Because the time between posing a scientific question and receiving the answer determines how fast progress in research is made, the challenge of turn-around time for big data analyses has to be addressed. As awareness of data sets and resources is essential, this has to be done in close collaboration with task areas 1 and 2.

Machine learning (ML) is a data analysis tool that has become indispensable for scientists of all disciplines. Applying ML in our research field is especially attractive because of the availability of high-quality labeled data from simulations. In addition, we can offer expertise on uncertainty treatments and on the development of solutions for large-scale and diverse collaborations. As part of the NFDI, we will build on this expertise by developing a service that provides automated ML for scientific data in the PAHN domain. This reduces the need for dedicated optimisation and training on many similar problems. The large variety of PAHN data sets ensures flexible and general solutions that can also support other areas of science. Task area 3 also includes the new aspect of tool preservation as ML methods are not necessarily reproducible from the documented network architecture.

Task area 4 ”Online analysis and data reduction”

Sophisticated data reduction and data compression methods are becoming increasingly important in order to cope with the huge amounts and rates of data. It is, however, extremely important not to lose any relevant information. This applies to all data-intensive sciences, but especially to the highly developed detectors in the PAHN facilites where methods must be developed that lead to drastic reductions in data volume already at the data source. This can only be achieved if competence in big data analytics is combined with the relevant scientific expertise. In addition, the increase of data volumes has tremendous impact on the long-term storage of scientific data; i.e. only a tiny fraction even of the reduced data can be stored. This requires a change in paradigm in almost any scientific sector, but in particular in astronomy and PAHN physics. All steps will be based on complex algorithms that have to select the relevant information out of enormous data streams.

In typical PAHN experiments, data are recorded as triggered events each containing a high dimension of raw data, further complemented by a still large number of physics-relevant features generated from the raw data. Many dimensions of this data set contain missing values or are even sparsely populated. Others will contain redundant or even no information relevant for a specific analysis. One aim of the efforts in machine learning based data reduction is to develop methods to identify a robust resource-optimised data set with maximal relevance and minimal redundancy containing all information needed to reproduce not only one analysis but a defined set of analyses. The focus of our activities in this task area is to provide generally valid and cross-disciplinary solutions for data reduction and data compression algorithms.

5.3 Governance

The foreseen consortium structure is displayed in figure 1). Based on the continuous feedback from the implemented task areas and cross-cutting topics about potentially changing requirements in the process of developing new data management infrastructures and services, this structure and corresponding management policies will be updated during the life cycle of the PAHN-PaN Consortium.

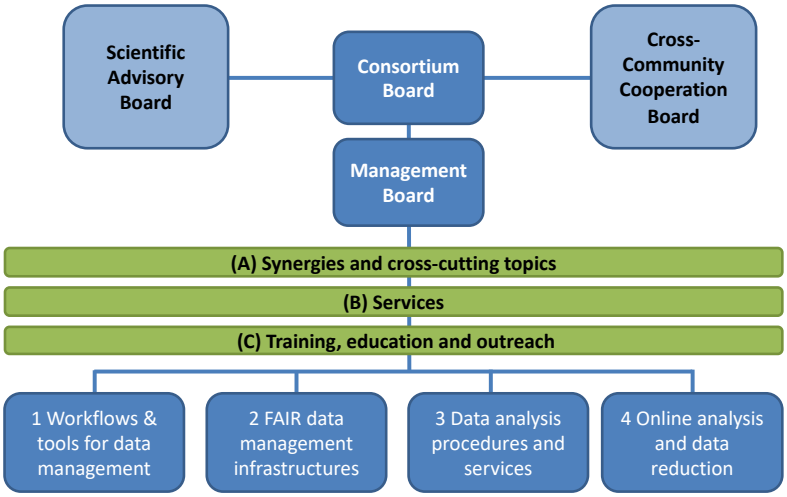

Figure 1: The foreseen governance structure of the PAHN-PaN Consortium.

The governing body of the PAHN-PaN Consortium will be the **Consortium Board (CB)** — a consortium parliament with one representative and one vote per applicant or co-applicant institution. The consortium participants have a right to be heard in the Consortium Board. The Consortium Board defines the objectives, strategy and structure of the consortium and selects the PAHN-PaN Project Manager and the other members of the Management Board as well as the task area leaders of task areas 1-4 and of the cross-cutting topics A-C.

In its tasks the CB is supported by a **Scientific Advisory Board (SAB)**, in which experts from the field of scientific computing in different communities discuss the overall scientific orientation of PAHN-PaN. In the SAB, care is taken to particularly represent the views of our international collaborators, keeping in mind that structural efforts in scientific computing can only be successful when agreed upon by all international partners from, e.g., the large experimental collaborations in PAHN physics. Also scientific librarians will be represented here.

The **Cross-Community Cooperation Board (CCCB)** is composed of members from PAHN-PaN and from other consortia. Its task is to make sure that PAHN-PaN and neighbouring consortia are in constant exchange and complement each other in their efforts. The CCCB in particular supervises and guides the concrete synergetic actions of the cross-cutting topics (A-C).

Day-to-day business of PAHN-PaN is handled by the **Management Board (MB)**, composed of one Project Manager (PM) and a number of board members. These are selected by the CB. The task of the MB is to organise the consortium internally and in its administrative affairs with the outside world. Concretely, the following tasks will be pursued here:

- financial and organisational management of the consortium;
- reporting and controlling;
- national and international interaction;
- interaction with industry on the political and administrative level;
- definition of policies and guidelines;
- reflection and integration of developments into the international collaborations;
- know-how transfer within consortium members and to/from NFDI partner consortia;
- execution of CB decisions.

The task areas and cross-cutting activities are organised by a number of scientists appointed by the CB.